PNAS
www.pnas.org

Supplementary Information for

Machine learning potentials for complex aqueous systems made simple

Christoph Schran, Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, Ondrej Marsalek and Angelos Michaelides

Christoph Schran and Angelos Michaelides.
E-mail: cs2121@cam.ac.uk, am452@cam.ac.uk

This PDF file includes:

Supplementary text
Figs. S1 to S7 (not allowed for Brief Reports)
SI References

Supporting Information Text

Quality assessment

In order to validate the six models developed by our rapid machine learning framework, we have established a validation protocol that makes it possible to compare the performance of the various models for different systems in a direct and condensed manner. For that purpose we have selected three main categories for structural and dynamical properties as well as the precision of the force prediction. The main idea behind this validation protocol is that it probes the performance of a given model for the thermodynamic condition it is developed for. Thus, all three categories are compared directly against the AIMD simulation that was used as the starting point of the development of the models. This enables the straight-forward evaluation of the models without the need for additional benchmark simulations with the expensive reference method.

Structural properties for complex liquid-solid systems are directly probed by the radial distribution functions (RDFs) of the various species, which provide detailed insight into the two-component structural arrangement. For the scoring of the models, we compute all RDFs for a given system, both for the AIMD reference simulation and for independent C-NNP simulations using the developed model. For an N component system this results in $\binom{N+1}{2}$ RDFs which can be directly compared between the AIMD and C-NNP results by using a suitable norm d^{RDF}

$$d^{\mathrm{RDF}} = 1 - \frac{\int_0^{+\infty} \left| g^{\mathrm{AIMD}}(r) - g^{\mathrm{C\text{-}NNP}}(r) \right| \mathrm{d}r}{\int_0^{+\infty} g^{\mathrm{AIMD}}(r)\mathrm{d}r + \int_0^{+\infty} g^{\mathrm{C\text{-}NNP}}(r)\mathrm{d}r} \qquad [1]$$

which provides a measure of the similarity of two RDFs ranging from 0 (for most different) to 1 (for identical). Averaging over all $\binom{N+1}{2}$ norms d^{RDF} finally yields a single number that can be converted into percent to provide a condensed score for the performance of the C-NNP model for structural properties.

Dynamical properties are directly encoded by the vibrational density of states (VDOS), which is obtained by Fourier transform of the velocity autocorrelation function. The VDOS can be computed separately for all components in a system of interest and thus provides detailed insight into the dynamical properties of the system, probing vastly different processes over a broad frequency range. For an N component system N species-resolved VDOS are computed, both for the AIMD and C-NNP simulations. Using a similar norm d^{VDOS} as for the RDFs

$$d^{\mathrm{VDOS}} = 1 - \frac{\int_0^{+\infty} \left| f^{\mathrm{AIMD}}(\nu) - f^{\mathrm{C\text{-}NNP}}(\nu) \right| \mathrm{d}\nu}{\int_0^{+\infty} f^{\mathrm{AIMD}}(\nu)\mathrm{d}\nu + \int_0^{+\infty} f^{\mathrm{C\text{-}NNP}}(\nu)\mathrm{d}\nu} \qquad [2]$$

the AIMD and C-NNP results can be condensed into a measure of the similarity between the different functions, which after averaging over the different species and conversion into percent provides the score of a C-NNP model for dynamical properties.

Finally, the precision of the C-NNP model for the prediction of the forces is evaluated to provide another property score. The forces are what drives the dynamics of the system of interest and are thus of fundamental importance for an accurate description. We generated a test set for the evaluation of the force performance, by selecting a large subset of 1000 structures and associated forces from the original AIMD simulation. The root mean square error (RMSE), calculated separately for the N species in each system

$$F^{\mathrm{RMSE}} = \sqrt{\frac{\sum_{i=1}^{3M} \left(F_i^{\mathrm{AIMD}} - F_i^{\mathrm{C\text{-}NNP}} \right)^2}{3M}} \qquad [3]$$

is used as a suitable measure for the force prediction. Since the magnitude of the forces can fluctuate strongly for different systems, but also within a given system (solid compared to liquid atoms), the RMSE is put into relation of the average force fluctuation of a given species

$$F^{\mathrm{RMS}} = \sqrt{\frac{\sum_{i=1}^{3M} \left(F_i^{\mathrm{AIMD}} \right)^2}{3M}}. \qquad [4]$$

The resulting N scaled force errors $F^{\mathrm{Force}} = \frac{F^{\mathrm{RMSE}}}{F^{\mathrm{RMS}}}$ are averaged and converted into percent to provide the force score of the C-NNP model.

Properties evaluated for Quality Assessment. All individual properties evaluated for the validation protocol comprising the final RDF, VDOS, and force score, as presented in the main text, are shown in full detail for all six C-NNP models in Fig. S1 to Fig. S6. Overall, essentially perfect agreement between the AIMD and C-NNP properties is observed for all six systems.

In addition, we have evaluated other properties for the systems of water under confinement or at interfaces in order to validate our C-NNP models in more detail. These properties are the density profiles, number of hydrogen bonds, and water orientation with respect to the interfaces for the CNT$-$H$_2$O, BNNT$-$H$_2$O, MoS$-$H$_2$O, and TiO$_2-$H$_2$O systems. They are all depict in Fig. S7 where overall substantial agreement between the C-NNP prediction and the much shorter AIMD reference simulations is observed. In particular, the density profile of the liquid and solid subsystems as shown in the upper two columns highlight the different nature of confinement with distinct density modulations that are all well reproduced by our C-NNP models. In addition, the number of hydrogen bonds along the water density profile is in substantial agreement between AIMD reference and C-NNP prediction for all four systems. Finally, the orientation of water with respect to the involved interfaces, as encoded by the cosine of the angle between the water dipole vector and the normal of the interface, is also well reproduced by our C-NNP models.

Computational Details

C-NNP models. All C-NNP models for the six selected systems were trained with the active learning workflow implemented in the AML Python package, available at https://github.com/MarsalekGroup/aml.

Using an *ab initio* trajectory as input, we construct a C-NNP model and its associated training set in an active learning protocol. 20 randomly selected structures from the trajectory are used to generate an initial C-NNP model. Next, the model is improved by iteratively adding structures with largest mean force committee disagreement to the training set, which is continued until convergence of the committee disagreement is observed. We performed 15 such active learning steps for all systems studied here, identifying 20 new structures for the training set in every step, but keeping out previously selected structures, which results in a total training set size of about 300 structures. Within every active learning step, the committee disagreement of 2000 randomly selected structures from the AIMD reference trajectory is evaluated and the 20 structures with largest mean force disagreement are added to the training set. The final C-NNP models are then obtained after stringent training of the NNP members with tight convergence criteria, as mentioned below in detail.

The chemical environment around each atom is described by a set of atom–centered symmetry functions (1), which transform the structure into suitable input for the atomic NNs. We applied a general set of symmetry functions to all systems studied here that can be automatically generated for a new system of interest within the AML package. The structural information for angular and radial symmetry functions is restricted to a radial cutoff of 12 bohr by a cosine cutoff function. For every pair of elements we employ ten radial symmetry functions, with fixed Gaussian width of 0.308 bohr, which are equally distributed within the 12 bohr cutoff. For every triple of elements we use four angular symmetry functions with a fixed Gaussian width of 0.012 bohr, $\lambda = \pm 1$ and ζ of 1 and 4, respectively. All symmetry function values are scaled and centered based on the average and range of the individual symmetry functions encountered in the training set according to

$$G^i = \frac{G^i - G^i_{\text{avg}}}{G^i_{\text{max}} - G^i_{\text{min}}}. \qquad [5]$$

All NNs consist of two hidden layers with 20 neurons, while the hyperbolic tangent was used as activation function for all layers, except the output neuron, which features a linear activation function. NNP optimizations are performed with the open-source n2p2 code (2) and the optimization parameters have been chosen according to the detailed benchmarking of this code for water (2).

Each C-NNP model is made up of 8 NNP members, which are constructed by random subsampling of the full reference data, where 10% of the points are left out in each case to impose the required diversity between C-NNP members. After different random initialization for each committee member, the weights and biases of the NNs were optimized using the parallel multistream version (2) of the adaptive global extended Kalman filter as implemented in n2p2. C-NNPs used for QbC were optimized for 15 epochs with 6 streams, while the final C-NNPs, to be used for simulations, were optimized for 50 epochs with 24 streams. All training input files, training sets and parameters of the final models are publicly available at https://github.com/water-ice-group/simple-MLP.

AIMD simulations. All *ab initio* molecular dynamics simulations used as input for our machine learning framework haven been performed with the CP2K software package (3).

The fluoride ion in water was described at the hybrid DFT level with the revPBE0 functional and D3 dispersion correction. The wavefunction was represented up to a plane wave cutoff of 400 Ry in conjunction with the TZV2P basis set and GTH pseudopotentials. The Hartree-Fock exchange calculation is speed up using the auxiliary density matrix methods as implemented in C2PK. A single fluoride ion was described in a periodic box of 64 water molecules with a cell size of 12.445 Å. The system was propagated in the NVT ensemble at 300 K with a molecular dynamics timestep of 0.5 fs. Equilbiration with a CSVR thermostat and 30 fs coupling constant was performed for 3 ps, while the temperature for the 50 ps production run was maintained with a CSVR thermostat and 1 ps coupling constant.

The sulfate ion in water was described with the BLYP functional in combination with the D3 dispersion correction. A plane wave cutoff of 280 Ry was used, while the molecularly optimized TZV2P atomic basis set was employed for all elements in combination with GTH pseudopotentials. The system contains a single sulfate ion and 64 water molecules in a 12.41 Å periodic box. This setup was simulated in the NVT ensemble with a time step of 0.5 fs, while the temperature was maintained with a CSVR thermostat with a 50 fs coupling constant. Equilibration was performed for 5 ps followed by a 30 ps production run. This simulation has been used in a previous study on the effects of polarization for the properties of the sulfate ion (4).

Simulations of water confined in carbon and hexagonal boron nitride nanotubes were performed for (12,12) armchair nanotubes with a length of 3 unit cells at a water density of 1.0 g/cm^3. This results in 288 wall atoms and 65 water molecules for the carbon nanotube and 68 water molecules for the hexagonal boron nitride nanotube. The PBE functional with D3 dispersion correction was used, in combination with GTH pseudopotentials, a 460 Ry plane wave cutoff and the DZVP molecularly optimised basis set. Deuterium masses were used for the hydrogen atoms and the molecular dynamics time step was set to 1.0 fs. Systems were pre-equilibrated for 5 ps at a temperature of 500 K using velocity rescaling, while keeping the positions of the atoms in the confining nanotubes fixed. Production runs in the NVT ensemble were then performed using Langevin dynamics, at a temperature of 330 K. Statistics were collected for approximately 130 ps for each system.

For water confined within molybdenum disulphide, the *ab initio* reference simulation consists of 109 water molecules confined by single layer MoS_2 sheets of 168 atoms in a 22.545, 22.314, 11.500 Å periodic box. The optB88-vdW functional was used, with GTH pseudopotentials, a 550 Ry plane wave cutoff, a relative cutoff of 60 Ry and the DZVP molecularly optimised basis set for all elements. Equilibration was performed in the NVT ensemble, with a Nosé-Hoover chain thermostats of length 5 to maintain a temperature of 500 K for 5 ps. This was followed by a second equilibration stage at 400 K for a further 5 ps. Final data collection was performed at 400 K over a 30 ps period.

Water on the rutile titanium dioxide (110) surface was described by a system consisting of 80 water molecules on four O-Ti-O trilayers in a periodic box of dimension 11.836, 12.9938, 42.00 Å. This results in a 1.5 nm thick water film on the four trilayers with additional 15 Å vacuum to separate the periodic images in z-direction. The optB88-vdW functional was used in combination with GTH pseudopotential, a 400 Ry plane wave cutoff and the DZVP molecularly optimised basis set for all elements. After equilibration, the system was propagated for 30 ps in the NVT ensemble at 300 K maintained by a Nosé-Hoover chain thermostat of length 4 with a coupling constant of 40 fs. A 1.0 fs timestep in combination with deuterium masses for hydrogen atoms was used and the atoms of the lowest O-Ti-O trilayer — not in contact with water — were kept fixed.

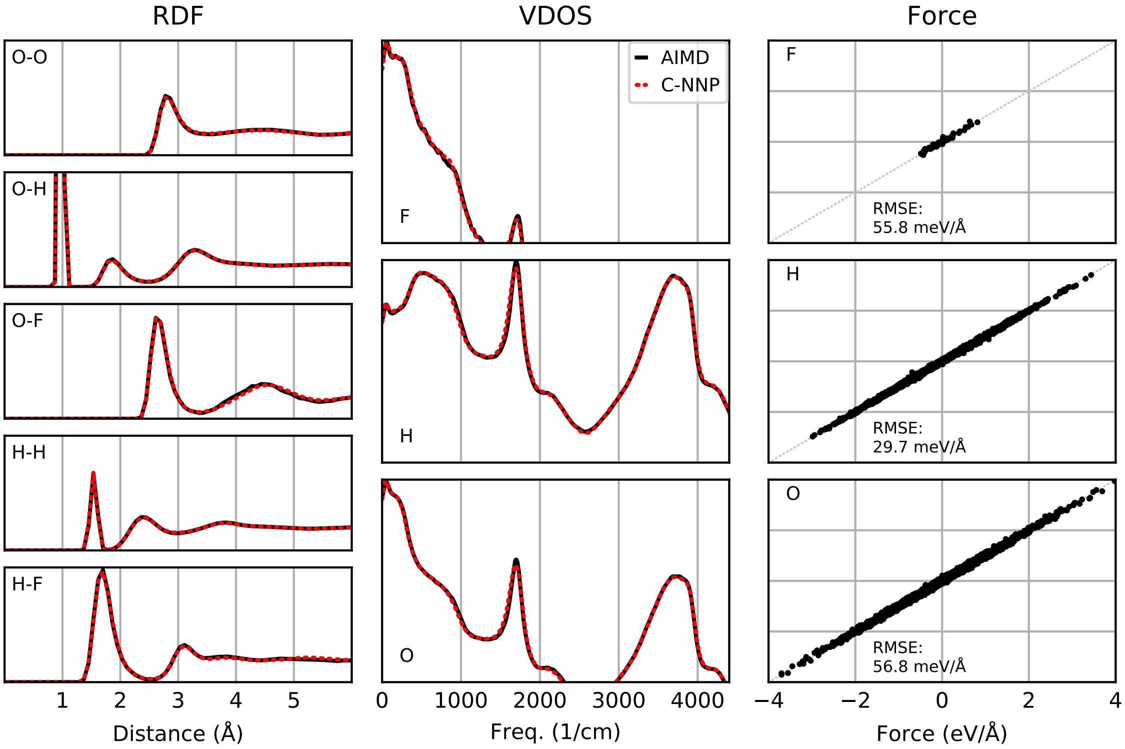

Fig. S1. Assessment of the precision of the fluoride-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

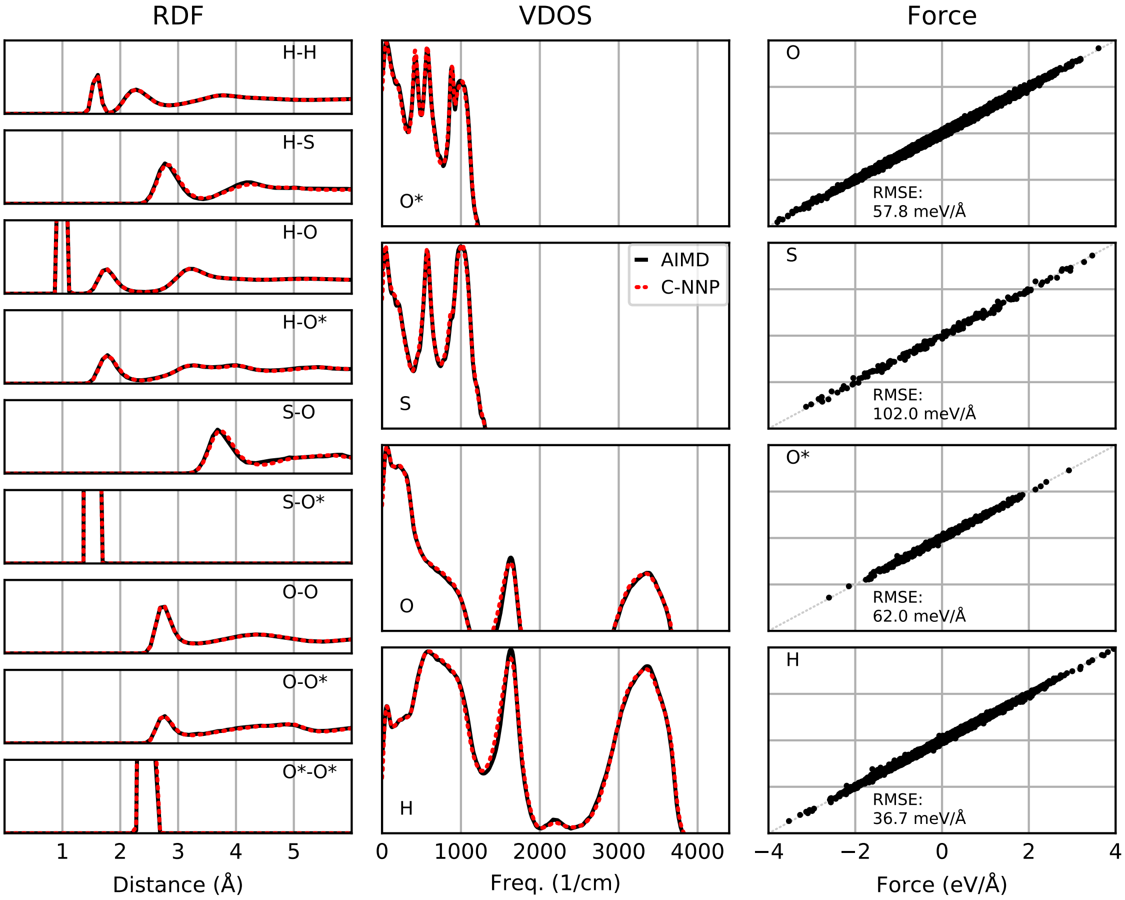

Fig. S2. Assessment of the precision of the sulfate-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

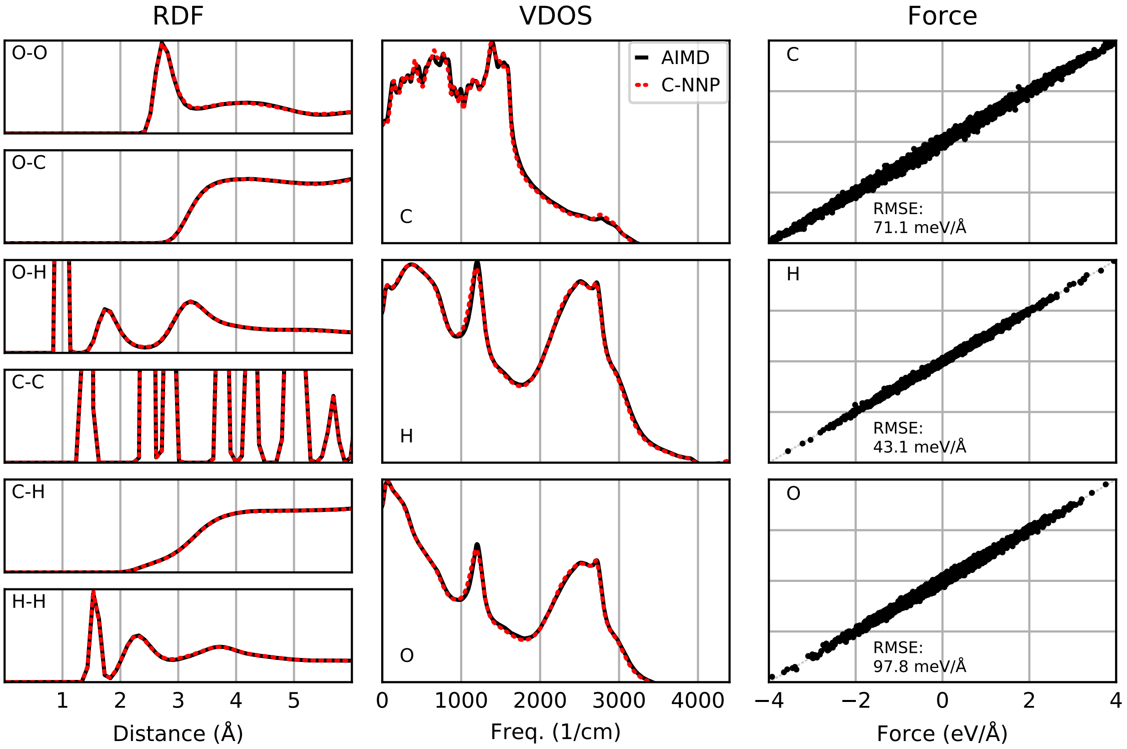

Fig. S3. Assessment of the precision of the carbon nanotube-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

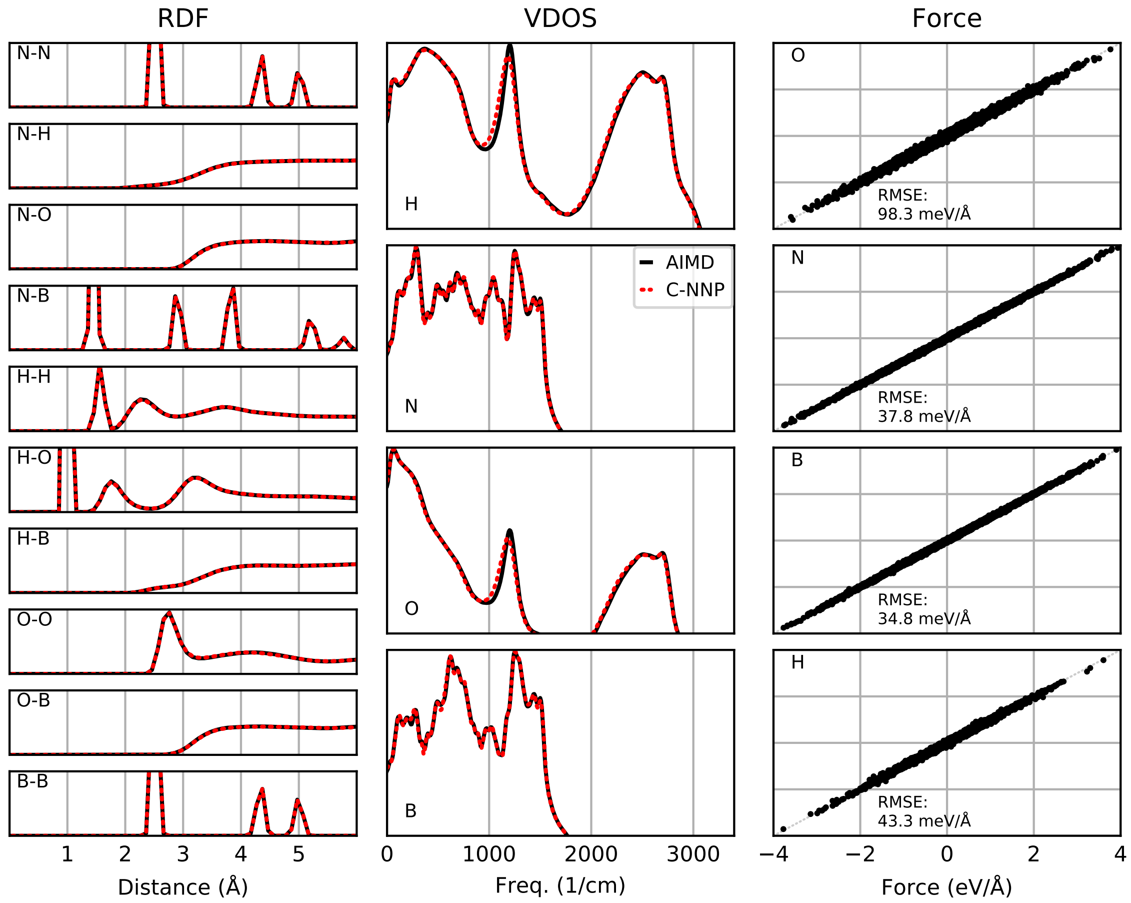

Fig. S4. Assessment of the precision of the hexagonal boron nitride nanotube-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

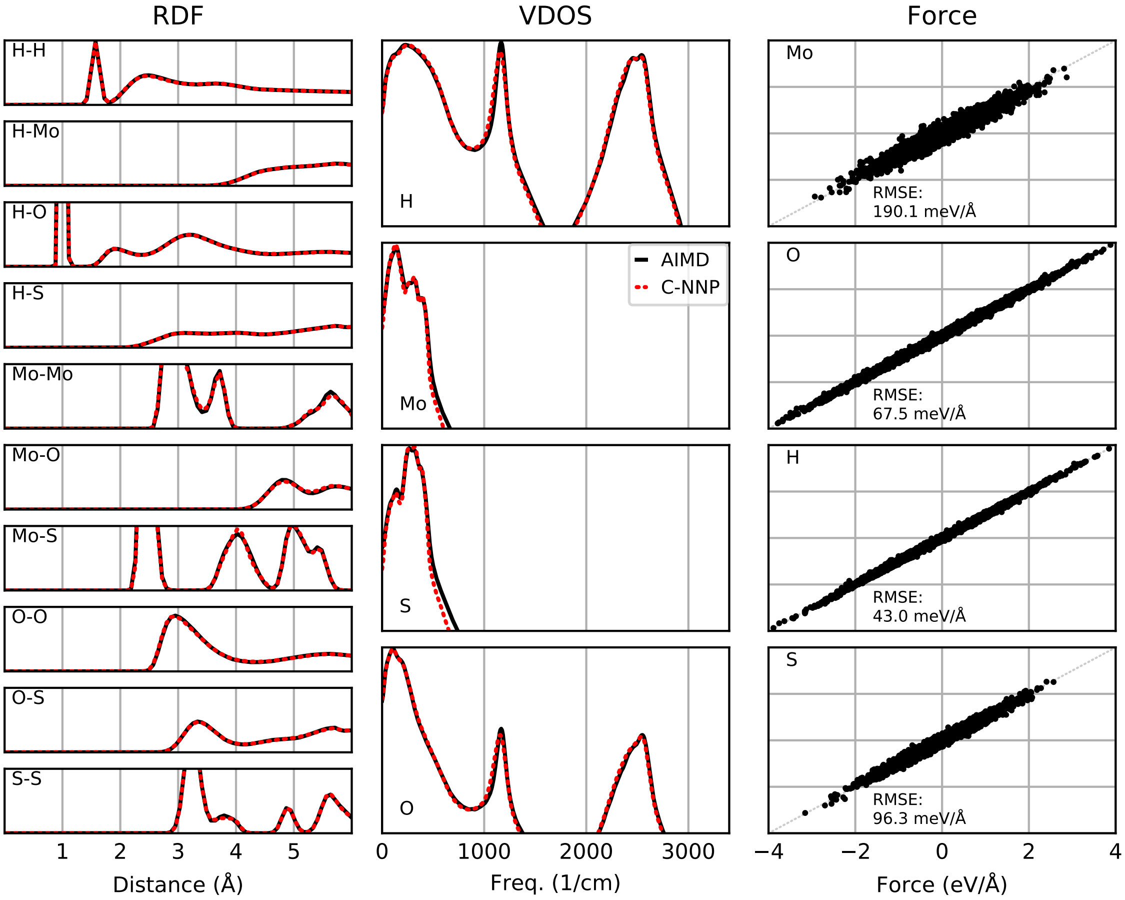

Fig. S5. Assessment of the precision of the molybdenum disulfide-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

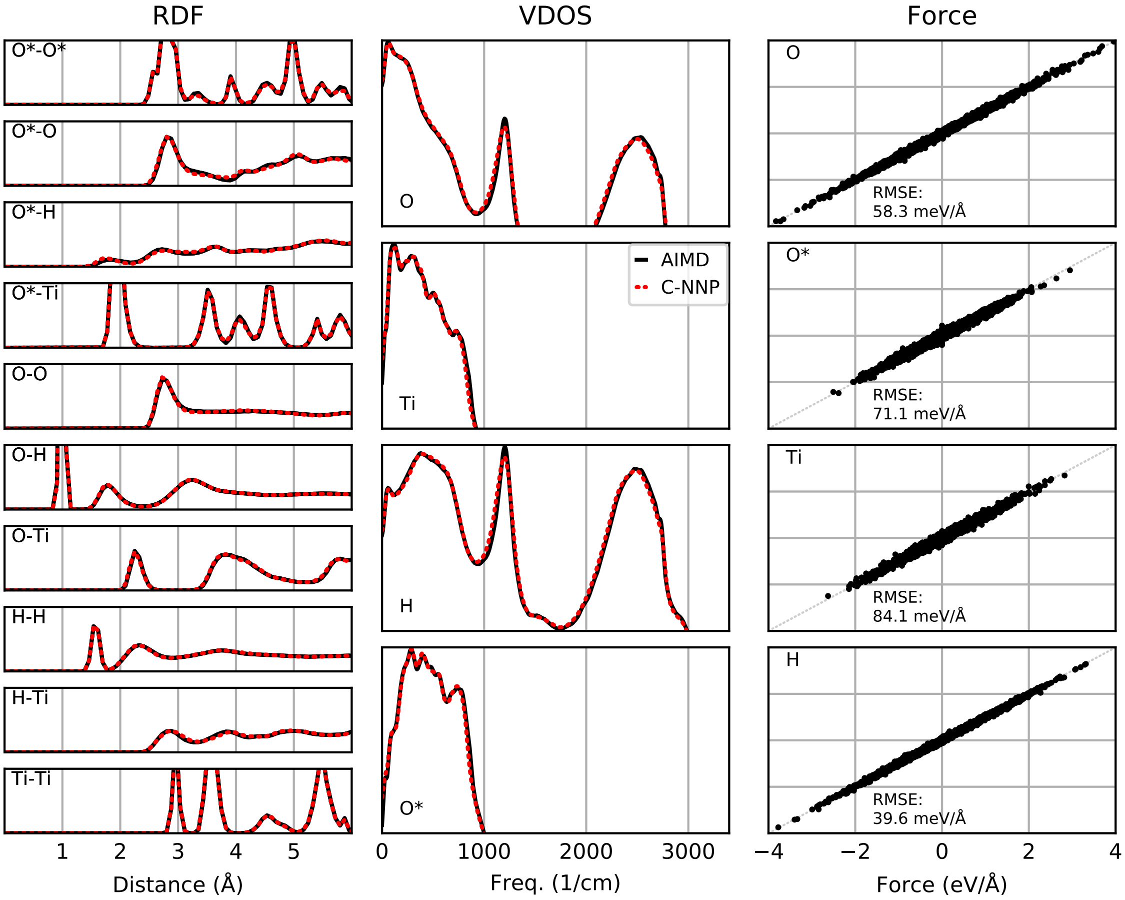

Fig. S6. Assessment of the precision of the titanium dioxide-water C-NNP model for structural and dynamical properties as well as the force prediction. The radial distribution functions (RDF) of all pairs of species are shown in the left panels comparing the AIMD and C-NNP results. The vibrational density of states (VDOS) of all species are shown in the middle panels comparing the AIMD and C-NNP results. The force correlation between AIMD and C-NNP forces of all species are shown in the right panels.

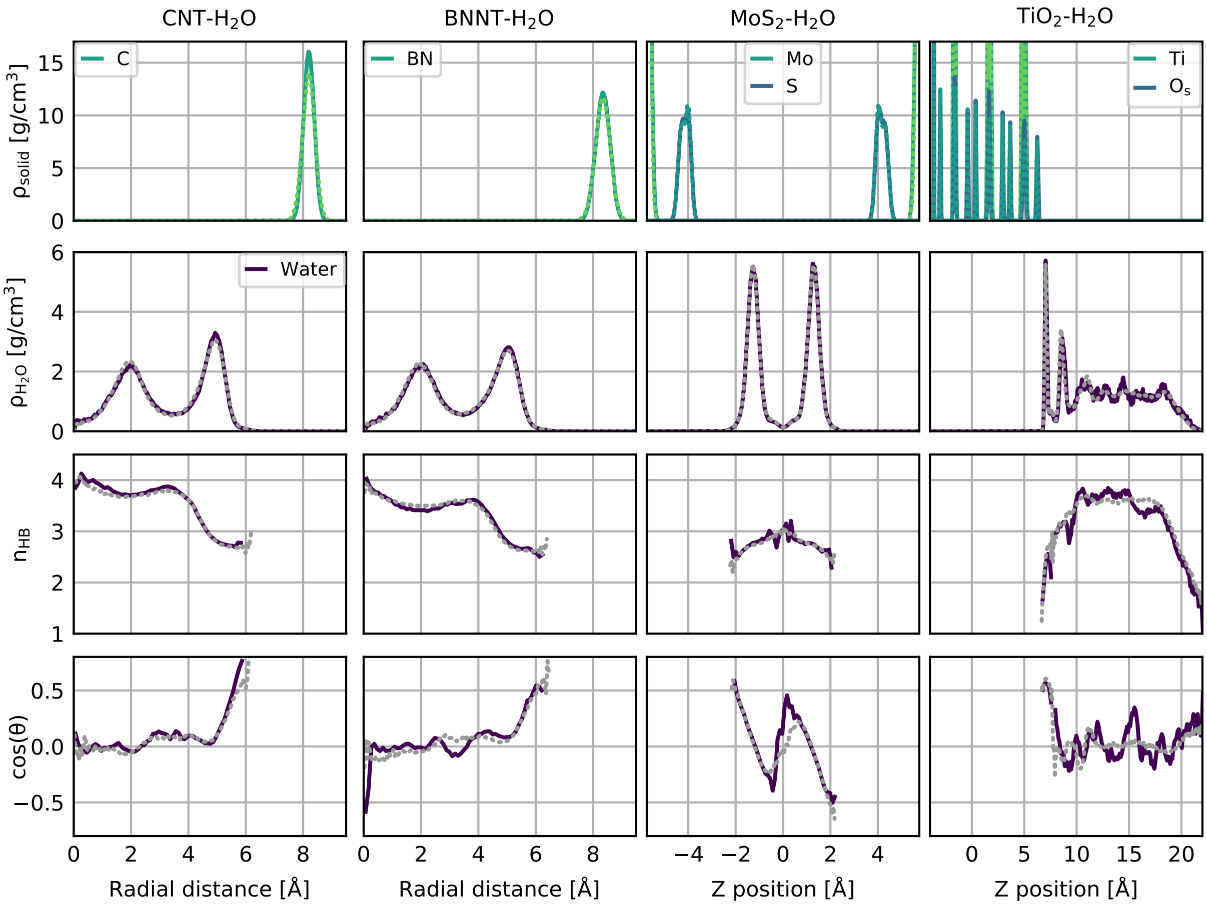

Fig. S7. Validation of system properties for four complex systems involving water under confinement or at interfaces. The first, second, third, and forth column compile results for water in a carbon nanotube (CNT—H_2O), water in a hexagonal boron nitride nanotube (BNNT—H_2O), water confined by single layer molybdenum disulfide (MoS_2—H_2O), and water at the rutile titanium dioxide surface (TiO_2—H_2O), respectively. The top row shows the density profiles of the involved solid subsystem and the second row the corresponding density of the water as a function of the radius for the two nanotubes and Z position for the others. The third row depicts the number of hydrogen bonds along the water density profile for the four systems, while the bottom row features the orientation of water with respect to the involved interface along the water density profile. The AIMD reference results are shown with solid lines, while the C-NNP predictions are included as dotted lines.

References

1. J Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
2. A Singraber, T Morawietz, J Behler, C Dellago, Parallel Multistream Training of High-Dimensional Neural Network Potentials. *J. Chem. Theory Comput.* **15**, 3075–3092 (2019).
3. TD Kühne, et al., CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 194103 (2020).
4. L Pegado, O Marsalek, P Jungwirth, E Wernersson, Solvation and ion-pairing properties of the aqueous sulfate anion: Explicit versus effective electronic polarization. *Phys. Chem. Chem. Phys.* **14**, 10248–10257 (2012).